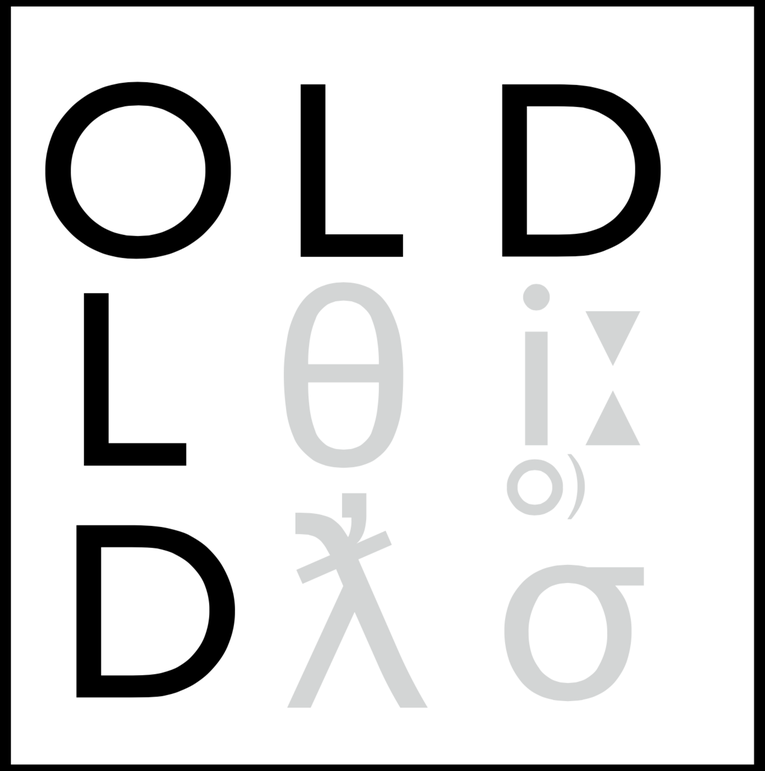
Dative:
from collaborative database to Archival Information Package

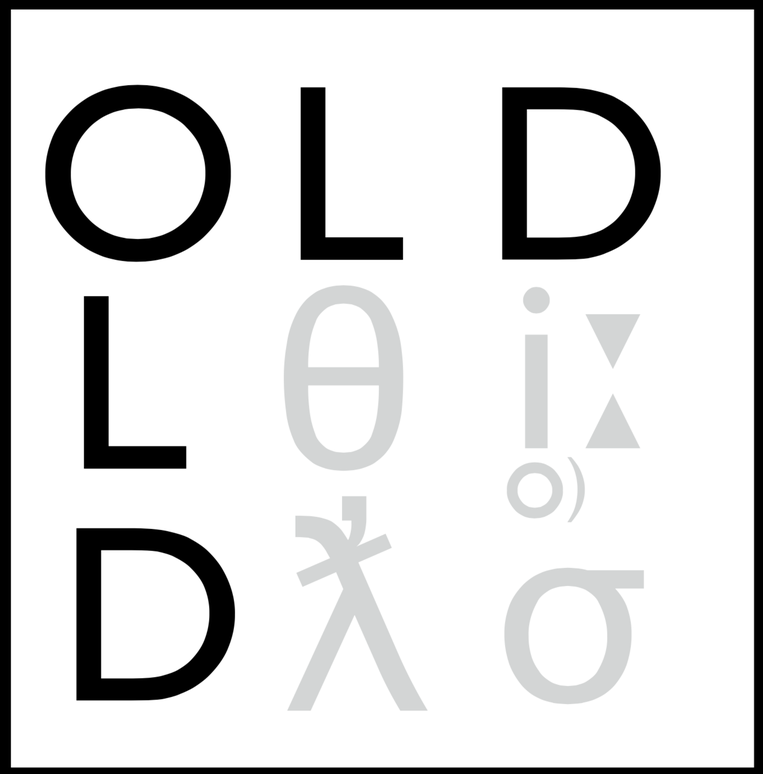
ICLDC 5
University of Hawai‘i at Mānoa
March 3, 2017

Joel Dunham (Concordia University & Artefactual Systems)



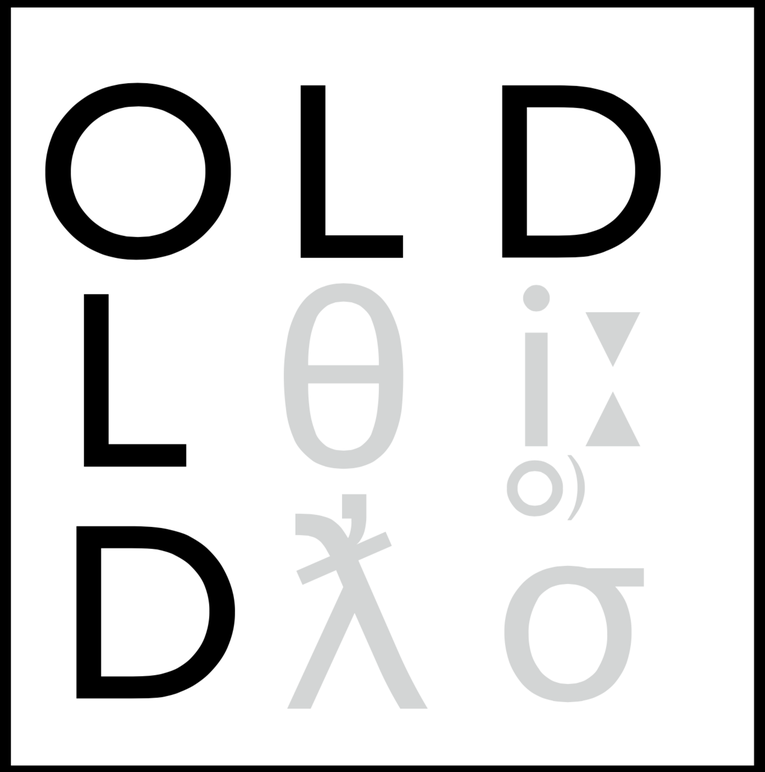
Open Source Linguistic Data Management

Dative: GUI



Online Linguistic Database

Archivematica: preservation pipeline

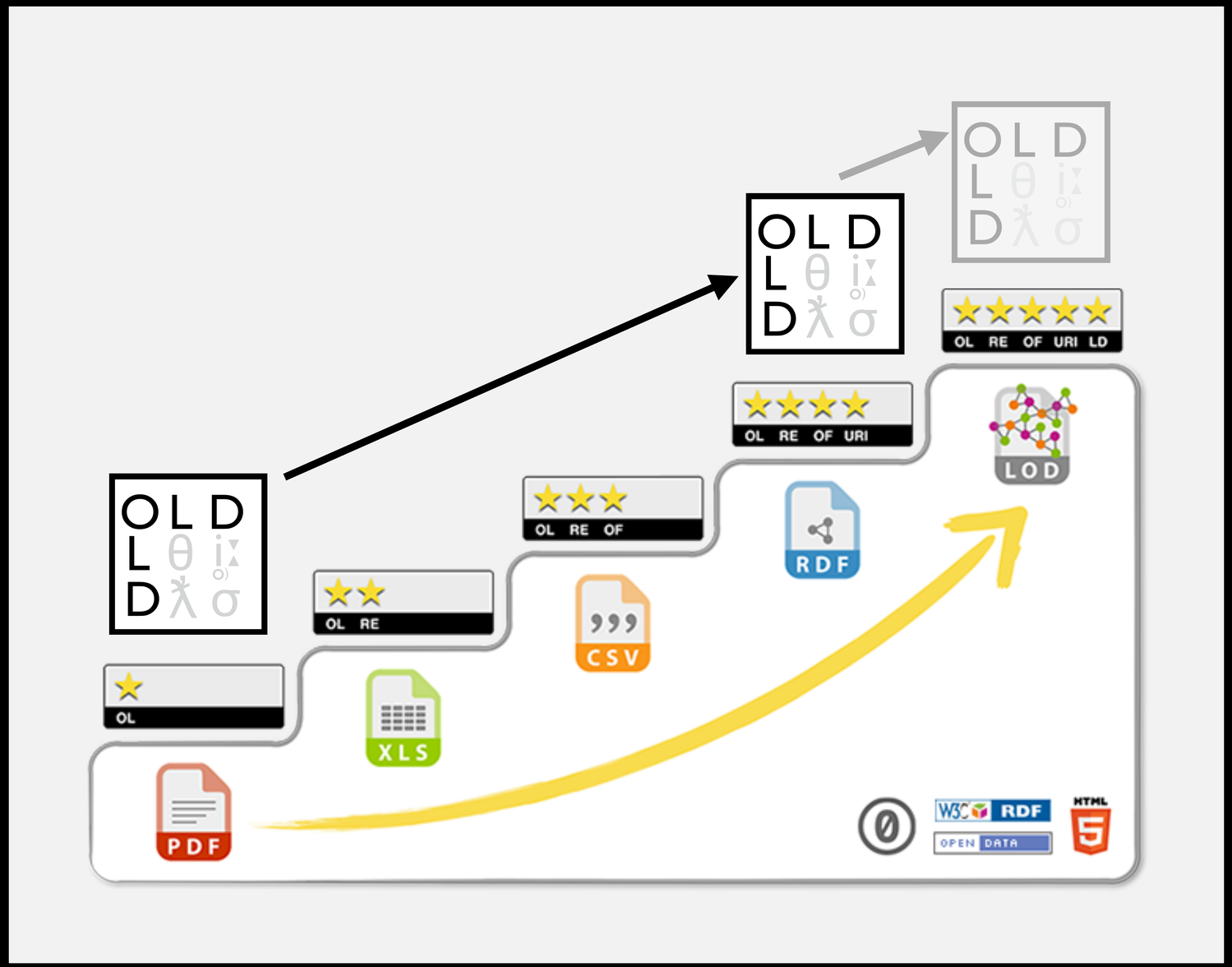


• manage • reuse • preserve

OLD LOD Export

- Linked Open Data (LOD) export
  - JSON-LD, Turtle, N-triples, RDF/XML
  - Uses existing ontologies: GOLD, Dublin Core, PROV, FOAF, Schema.org, SKOS, etc.
  - Static, citeable, reusable, web-available, semantically specified data sets

Towards 5 ★ Open Data



Existing Ontologies

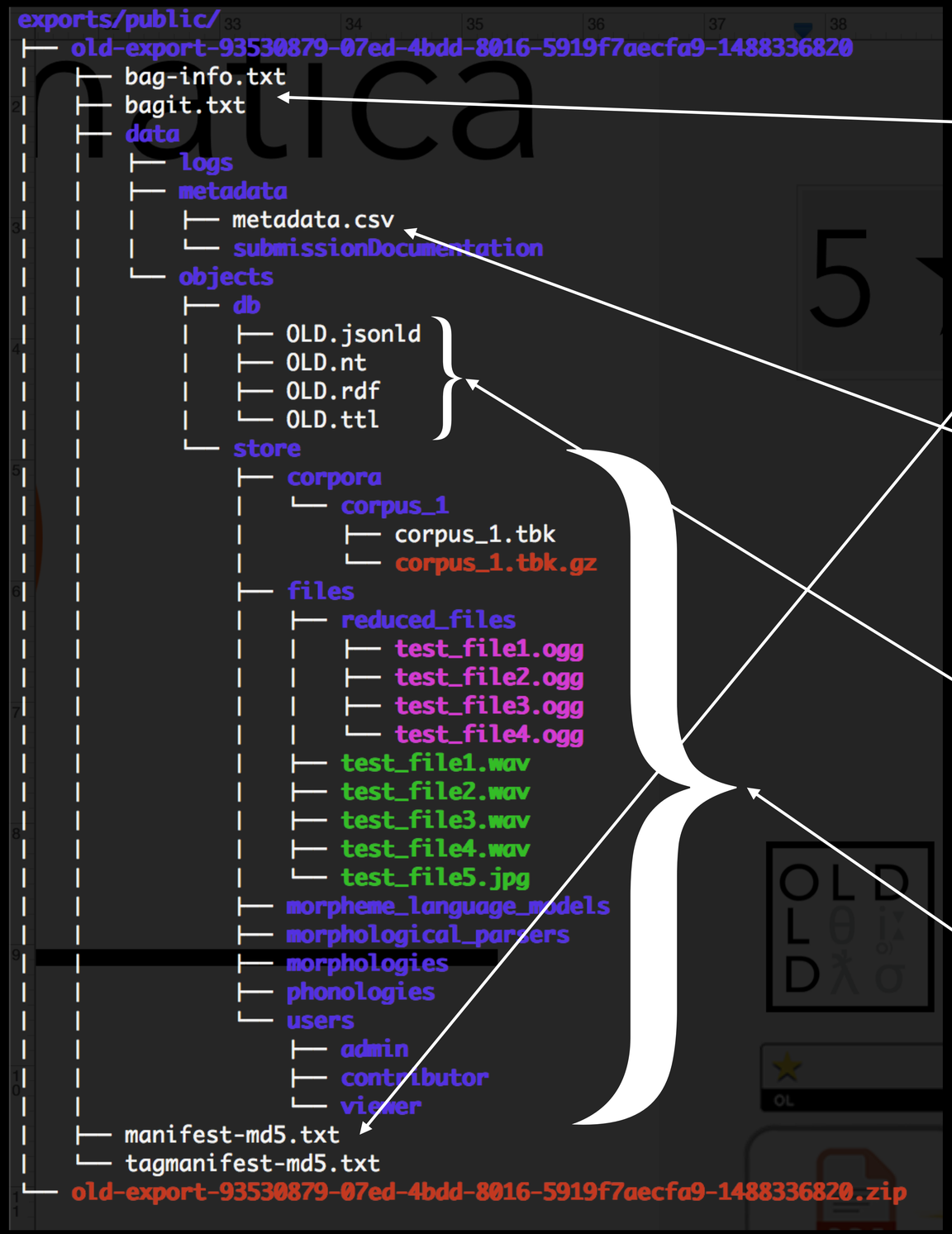
gold: General Ontology for Linguistic Description

lexvo: SIL's ISO 639-3 language codes

prov: provenance information: elicitation events, data entry events, etc.

sioc: Semantically Interlinked Communities: OLD accounts, users, roles, etc.

```
"gold": "http://purl.org/linguistics/gold/",
"asit": "http://ims.dei.unipd.it/websites/ASIt/RDF/doc/#",
"lexvo": "http://www.lexvo.org/page/iso639-3/",
"dc": "http://purl.org/dc/elements/1.1/",
"dcterms": "http://purl.org/dc/terms/",
"iana": "http://www.iana.org/assignments/media-types/",
"foaf": "http://xmlns.com/foaf/0.1/",
"schema": "http://schema.org/",
"skos": "http://www.w3.org/2004/02/skos/core#",
"pcdm": "http://pcdm.org/models#",
"org": "http://www.w3.org/ns/org#",
"dcat": "http://www.w3.org/ns/dcat#",
"muto": "http://purl.org/muto/core#",
"prov": "http://www.w3.org/ns/prov#",
"sioc": "http://rdfs.org/sioc/ns#",
```



BagIt (LoC) transmission format: structure, metadata & checksums

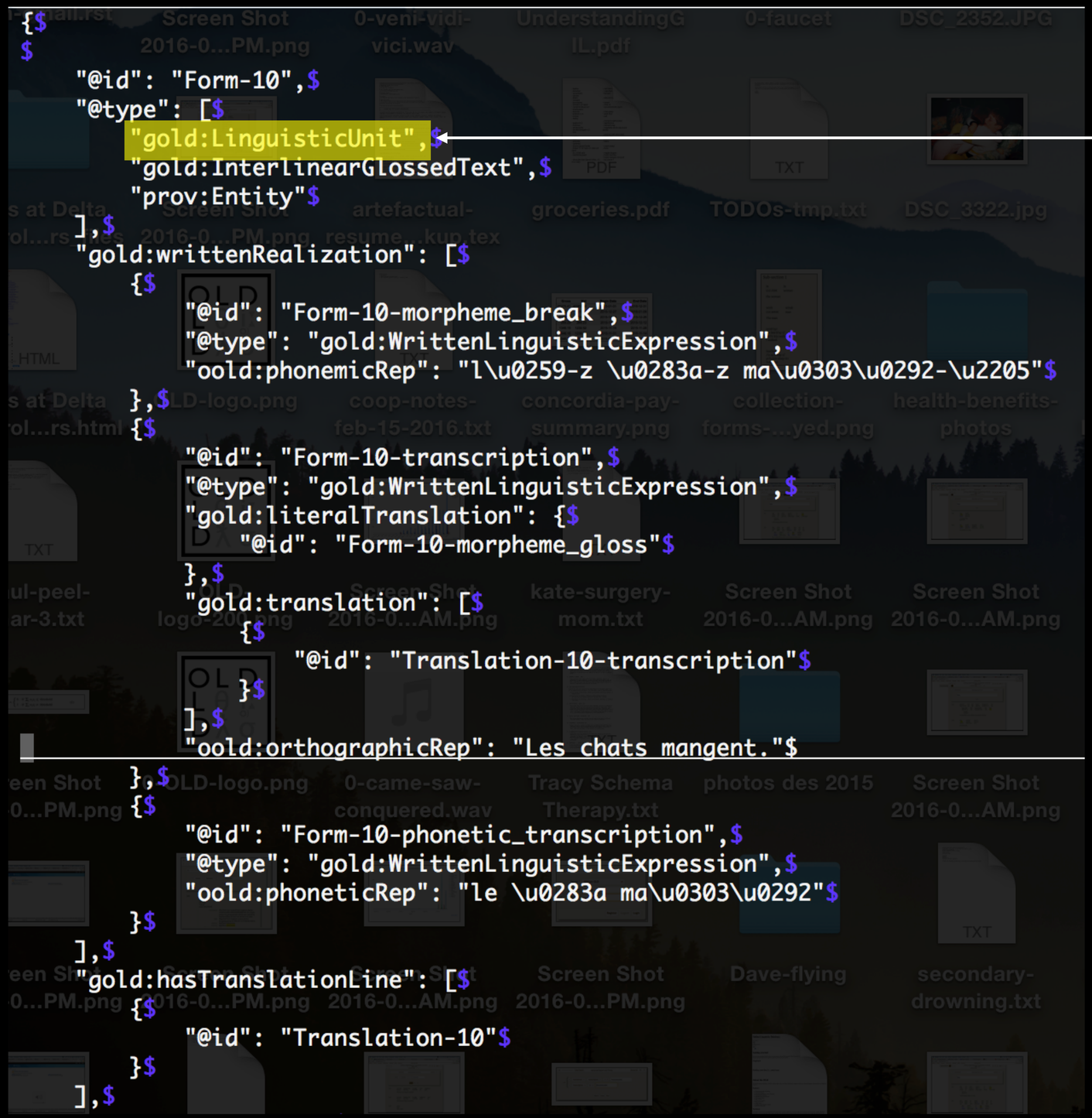
Dublin Core descriptive metadata about data set

OLD database as linked data (triples, JSON-LD, Turtle, etc.)

OLD binaries: audio/video, images, parsers, treebanks, etc.

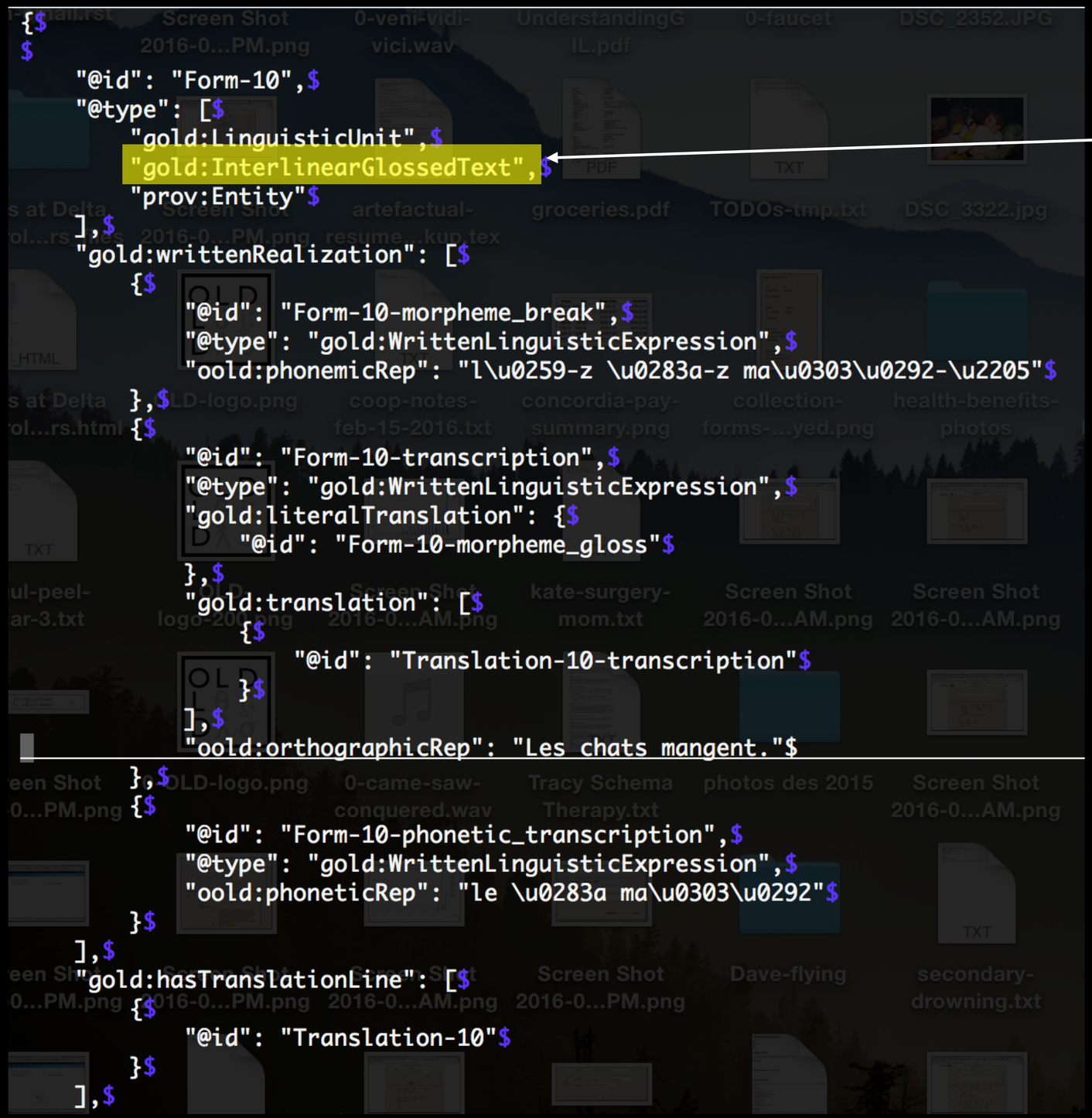
Why LOD Export?

- Data sets that make sense
  - even without Dative/OLD
    - even centuries from now
- Exit strategy: get your data out of Dative/OLD
- Reuse: others can query (SPARQL), harvest, link to, cite/attribute your public data sets



gold:LinguisticUnit

“This term is often used in linguistics and phonetics to refer to any entity which constitutes the focus of an enquiry. The unit is the stretch of language that carries grammatical patterns, and within which grammatical choices are made. For example, the unit sentence consists of one or more instances of the unit clause, and so on. [Crystal 2008: 503]”



```
{$
$
    "@id": "Form-10",$
    "@type": [$
        "gold:LinguisticUnit",$
        "gold:InterlinearGlossedText",$
        "prov:Entity"$
    ],$
    "gold:writtenRealization": [$
        {$
            "@id": "Form-10-morpheme_break",$
            "@type": "gold:WrittenLinguisticExpression",$
            "oold:phonemicRep": "l\u0259-z \u0283a-z ma\u0303\u0292-\u2205"$
        },$
        {$
            "@id": "Form-10-transcription",$
            "@type": "gold:WrittenLinguisticExpression",$
            "gold:literalTranslation": {$
                "@id": "Form-10-morpheme_gloss"$
            },$
            "gold:translation": [$
                {$
                    "@id": "Translation-10-transcription"$
                }$
            ],$
            "oold:orthographicRep": "Les chats mangent."$
        },$
        {$
            "@id": "Form-10-phonetic_transcription",$
            "@type": "gold:WrittenLinguisticExpression",$
            "oold:phoneticRep": "le \u0283a ma\u0303\u0292"$
        }$
    ],$
    "gold:hasTranslationLine": [$
        {$
            "@id": "Translation-10"$
        }$
    ],$
```

gold:InterlinearGlossedText

“Interlinear glossed text (IGT) is a linguistic data structure meant to display morphosyntactic structure: morphemes, morpheme boundaries, morpheme types (clitics, prefixes, reduplicated forms), morphosyntactic features/values and part of speech information. At a minimum, an instance of IGT includes a single line of source language followed by a translation line. […] Standards for IGT include the Leipzig Glossing Rules.”

```
{$
$
    "@id": "Form-10",$
    "@type": [$
        "gold:LinguisticUnit",$
        "gold:InterlinearGlossedText",$
        "prov:Entity"$
    ],$
    "gold:writtenRealization": [$
        {$
            "@id": "Form-10-morpheme_break",$
            "@type": "gold:WrittenLinguisticExpression",$
            "oold:phonemicRep": "l\u0259-z \u0283a-z ma\u0303\u0292-\u2205"$
        },$
        {$
            "@id": "Form-10-transcription",$
            "@type": "gold:WrittenLinguisticExpression",$
            "gold:literalTranslation": {$
                "@id": "Form-10-morpheme_gloss"$
            },$
            "gold:translation": [$
                {$
                    "@id": "Translation-10-transcription"$
                }$
            ],$
            "oold:orthographicRep": "Les chats mangent."$
        },$
        {$
            "@id": "Form-10-phonetic_transcription",$
            "@type": "gold:WrittenLinguisticExpression",$
            "oold:phoneticRep": "le \u0283a ma\u0303\u0292"$
        }$
    ],$
    "gold:hasTranslationLine": [$
        {$
            "@id": "Translation-10"$
        }$
    ],$
```

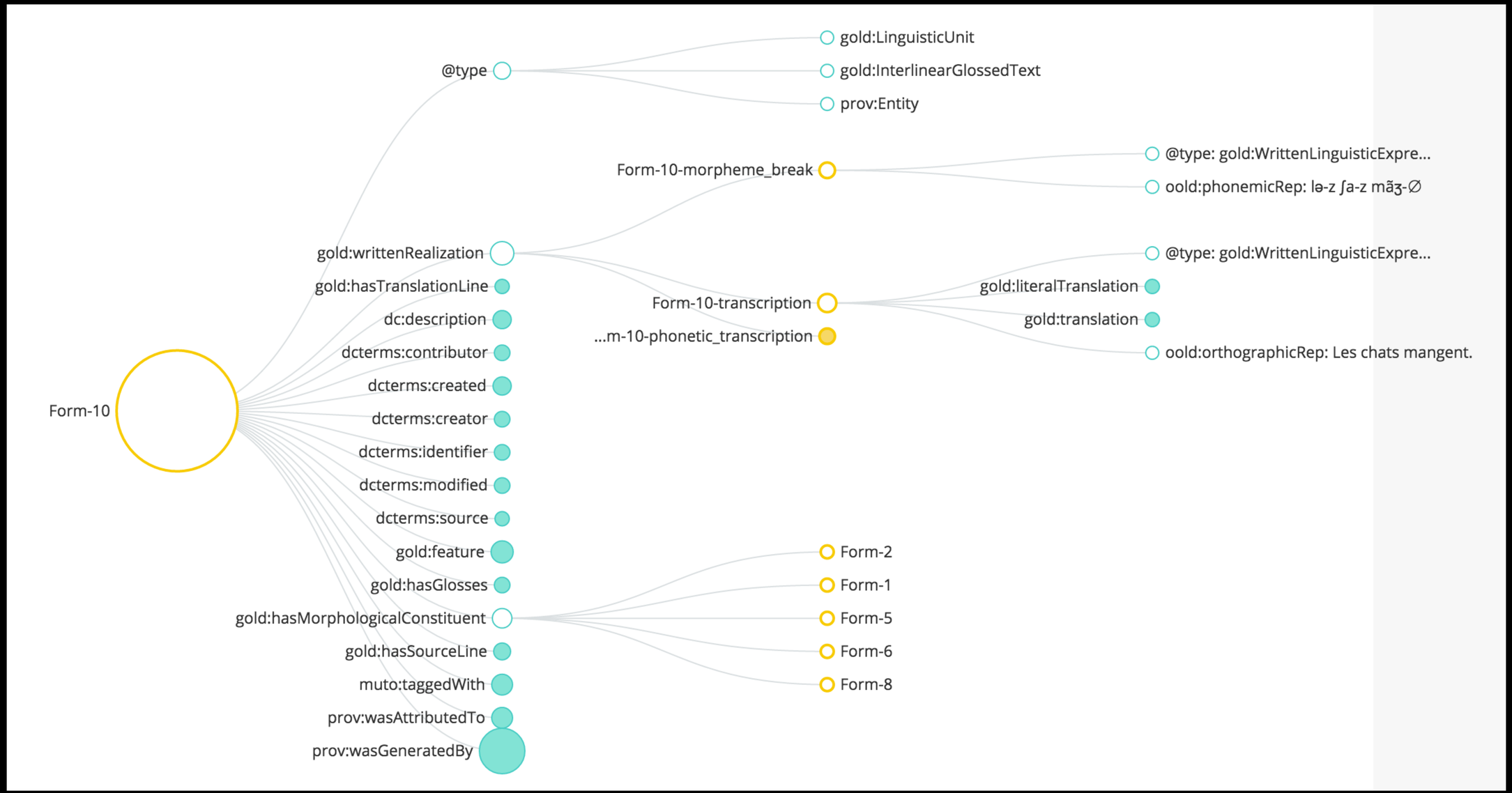
gold:writtenRealization

"The relation between some linguistic unit and its corresponding written expression."

gold:WrittenLinguisticExpression

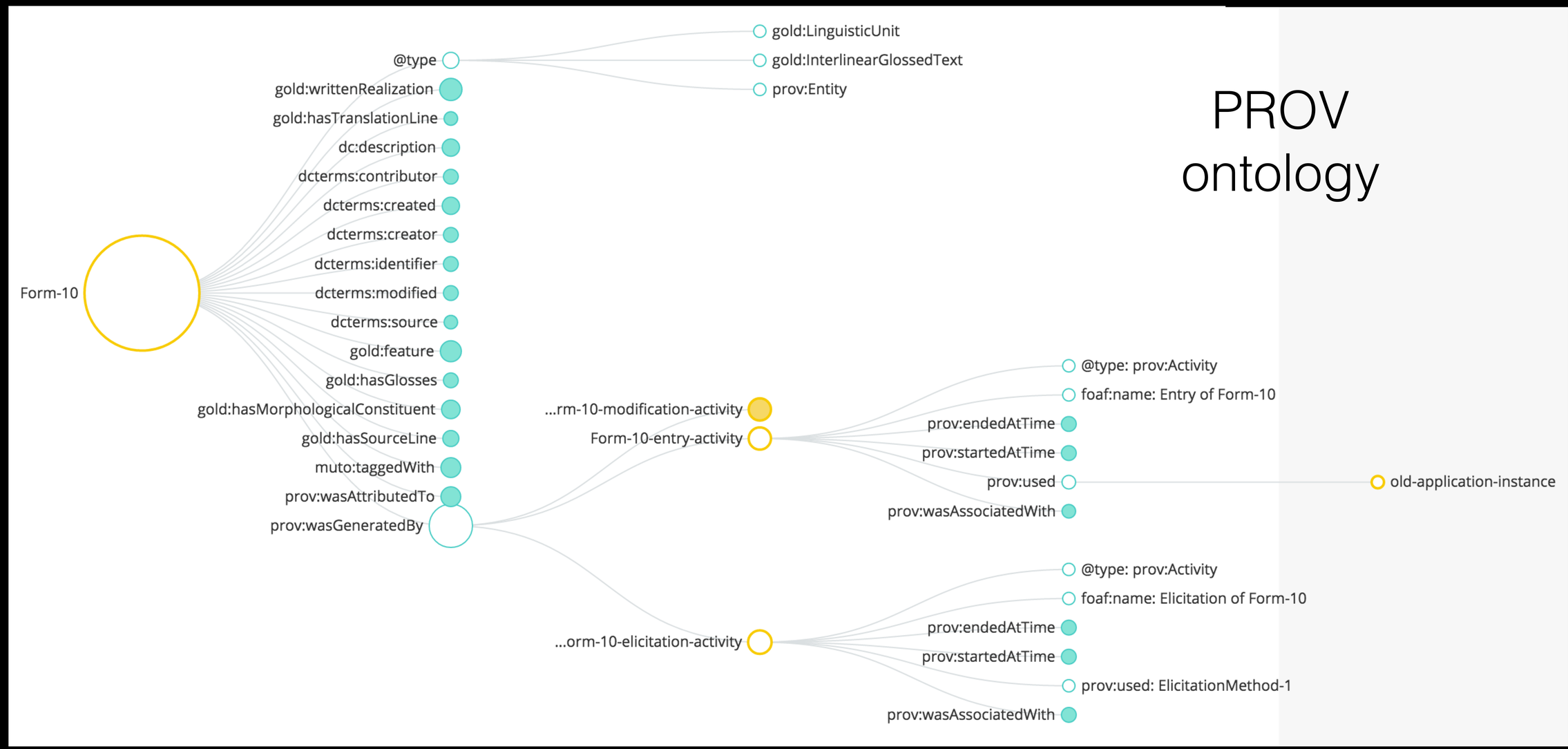
"The written physical form of language as distinct from either signed or spoken expressions. A written expression is the physical product of the writing process.."

OLD-independent linguistic description



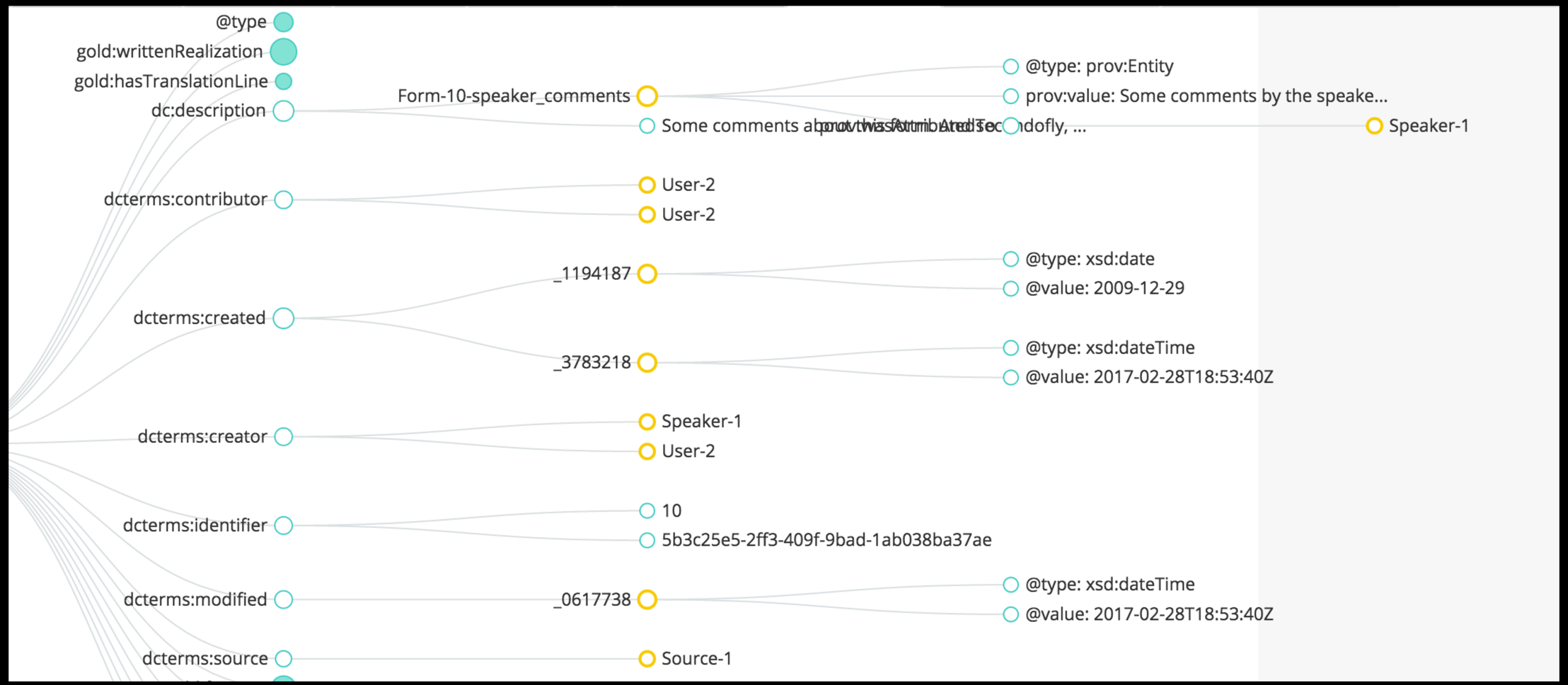
Provenance metadata

"This linguistic unit was created by means of an elicitation activity involving a speaker and an elicitor and an entry activity into an OLD application."

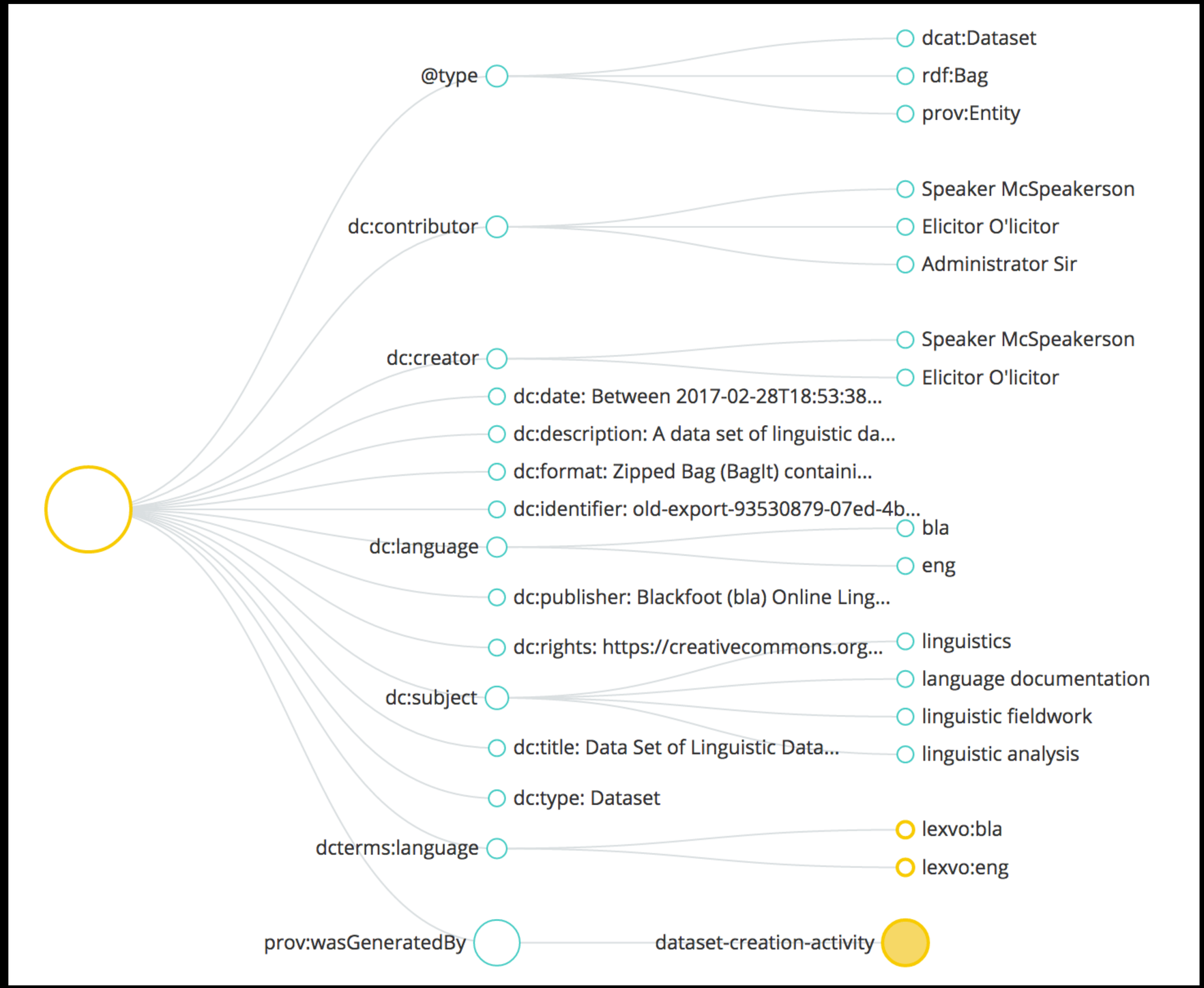


PROV ontology

Descriptive metadata (Dublin Core)



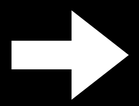
Descriptive metadata about entire data set



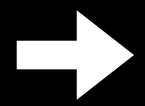
OLD LOD/BagIt Exports

- public or private
- public accessible at stable URLs
- queryable vai SPARQL / RDF libaries

Archivematica
is a preservation pipeline



- standard packaging
- metadata (METS, PREMIS)
- fixity
- file identification & characterizatic
- file normalization



data set

```
├── bag-info.txt
├── bagit.txt
├── data
│   ├── logs
│   ├── metadata
│   │   ├── metadata.csv
│   │   └── submissionDocumentation
│   └── objects
│       ├── db
│       │   ├── OLD.jsonld
│       │   ├── OLD.nt
│       │   ├── OLD.rdf
│       │   └── OLD.ttl
```

METS file: everything that happened to your data in the preservation process

Archival Information Package

```
├── bag-info.txt
├── bagit.txt
├── data
│   ├── METS.0d74e800-3854-41d3-be45-fa6184e433b5.xml
│   ├── logs
│   │   ├── fileFormatIdentification.log
│   │   ├── filenameCleanup.log
│   │   └── transfers
│   │       └── old-export-93530879-07ed-4bdd-8016-5919f7aecfa9-1488336820
│   │           └── logs
│   │               ├── BagIt
│   │               │   ├── bag-info.txt
│   │               │   ├── bagit.txt
│   │               │   └── tagmanifest-md5.txt
│   │               ├── extractContents.log
│   │               ├── fileFormatIdentification.log
│   │               └── filenameCleanup.log
│   └── objects
│       ├── db
│       │   ├── OLD.jsonld
│       │   ├── OLD.nt
│       │   ├── OLD.rdf
│       │   └── OLD.ttl
```

Archivematica METS / PREMIS

- Archivematica uses archival standards—PREMIS in METS—to encode technical & descriptive metadata about the OLD data set and the preservation actions taken on it.

- METS = Metadata Encoding and Transmission Standard

- PREMIS = Preservation Metadata: Implementation Strategies

Descriptive Dublin Core metadata

from our OLD export →

```
<?xml version='1.0' encoding='ASCII'?>$
<mets:mets xmlns:mets="http://www.loc.gov/METS/" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLS
  <mets:metsHdr CREATEDATE="2017-02-28T19:04:03"/>$
  <mets:dmdSec ID="dmdSec_1">$
    <mets:mdWrap MDTYPE="DC">$
      <mets:xmlData>$
        <dcterms:dublincore xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.org/dc/terms/" xsi:schemaLocati
          <dc:contributor>first_name 1 last_name 1, Contributor Contributor, Admin Admin</dc:contributor>$
          <dc:creator>first_name 1 last_name 1, Contributor Contributor</dc:creator>$
          <dc:publisher>Blackfoot (bla) Online Linguistic Database instance (running OLD version 2.0.0) at URL https://app.onlinel
          <dc:date>Between 2017-02-28T18:53:38Z and 2017-02-28T18:53:40Z.</dc:date>$
          <dc:description>A data set of linguistic data on language Blackfoot (bla). This data set was created using the Online Li
          <dc:format>Zipped Bag (BagIt) containing JSON-LD.</dc:format>$
          <dc:identifier>old-export-93530879-07ed-4bdd-8016-5919f7aecfa9-1488336820</dc:identifier>$
          <dc:language>bla, eng</dc:language>$
          <dc:relation></dc:relation>$
          <dc:coverage></dc:coverage>$
          <dc:rights>https://creativecommons.org/licenses/by-sa/3.0/legalcode</dc:rights>$
          <dc:subject>linguistics, language documentation, linguistic fieldwork, linguistic analysis</dc:subject>$
          <dc:title>Data Set of Linguistic Data on Language Blackfoot (bla) (created by the Online Linguistic Database)</dc:title>
          <dc:type>Dataset</dc:type>$
        </dcterms:dublincore>$
      </mets:xmlData>$
    </mets:mdWrap>$
  </mets:dmdSec>$
```

Technical metadata

checksum

PRONOM format id

MIME type

```
<mets:techMD ID="techMD_4">$
  <mets:mdWrap MDTYPE="PREMIS:OBJECT">$
    <mets:xmlData>$
      <premis:object xmlns:premis="info:lc/xmlns/premis-v2" xsi:type="premis:file" xsi:schemaLocati
        <premis:objectIdentifier>$
          <premis:objectIdentifierType>UUID</premis:objectIdentifierType>$
          <premis:objectIdentifierValue>bf958c51-e109-4f7e-b0ea-a446cbe820da</premis:objectIdentifi
        </premis:objectIdentifier>$
        <premis:objectCharacteristics>$
          <premis:compositionLevel>0</premis:compositionLevel>$
        <premis:objectCharacteristics>$
          <premis:compositionLevel>0</premis:compositionLevel>$
          <premis:fixity>$
            <premis:messageDigestAlgorithm>sha256</premis:messageDigestAlgorithm>$
            <premis:messageDigest>f2b6b25020b9c507b72c8dcd309632fa1dc4bc5d7c6504cd4b08cd38c53a94b0
          </premis:fixity>$
          <premis:size>183193</premis:size>$
          <premis:format>$
            <premis:formatDesignation>$
              <premis:formatName>Turtle</premis:formatName>$
              <premis:formatVersion>None</premis:formatVersion>$
            </premis:formatDesignation>$
            <premis:formatRegistry>$
              <premis:formatRegistryName>PRONOM</premis:formatRegistryName>$
              <premis:formatRegistryKey>fmt/874</premis:formatRegistryKey>$
            </premis:formatRegistry>$
          </premis:format>$
          <premis:objectCharacteristicsExtension>$
            <fits xmlns="http://hul.harvard.edu/ois/xml/ns/fits/fits_output" xsi:schemaLocation="ht
              <identification status="SINGLE_RESULT">$
                <identity format="Plain text" mimetype="text/plain" toolname="FITS" toolversion="0
                  <tool toolname="file utility" toolversion="5.14"/>$
                </identity>$
```

Digital provenance metadata

Archivematica created a .tif preservation derivative (normalization event) from an OLD .jpg file

```
<mets:digiprovMD ID="digiprovMD_156">$
  <mets:mdWrap MDTYPE="PREMIS:EVENT">$
    <mets:xmlData>$
      <premis:event xmlns:premis="info:lc/xmlns/premis-v2" xsi:schemaLocation="info:lc/xmlns/premis-v2 http://www.loc.gov/standa
        <premis:eventIdentifier>$
          <premis:eventIdentifierType>UUID</premis:eventIdentifierType>$
          <premis:eventIdentifierValue>1933b4f0-acc0-4ada-b5d5-d3cec484cb24</premis:eventIdentifierValue>$
        </premis:eventIdentifier>$
        <premis:eventType>normalization</premis:eventType>$
        <premis:eventDateTime>2017-02-28T19:02:22+00:00</premis:eventDateTime>$
        <premis:eventDetail>ArchivematicaFPRCommandID="a34ddc9b-c922-4bb6-8037-bbe713332175"; program="convert"; version="Versio
        <premis:eventOutcomeInformation>$
          <premis:eventOutcome></premis:eventOutcome>$
          <premis:eventOutcomeDetail>$
            <premis:eventOutcomeDetailNote>%SIPDirectory%objects/store/files/test_file5-a5ba16bc-b1e3-45ad-ba2a-5a5e90a265d0.tif-
          </premis:eventOutcomeDetail>$
        </premis:eventOutcomeInformation>$
```

- linked open data
- citable
- web-accessible
- reusable

- preservation-ready

Open Source

https://github.com/jrwdunham/dative

https://github.com/jrwdunham/old-pyramid

https://github.com/jrwdunham/deploy-dative-old

https://github.com/artefactual/archivematica

How to use Dative/OLD

A. Concordia University community server (free, up to 1GB storage space)

B. Install it yourself—it's open source and free (and has documentation and Ansible install scripts)

C. Contact LambdaBar (joel@lambdabar.com) for custom installations and feature development

Contribute to OLD's LOD semantics

- https://github.com/jrwdunham/old-pyramid/tree/dev/issue-jsonld-export